# SPECIFICATION OF THE 1996 HUB 4 BROADCAST NEWS EVALUATION

*Richard M. Stern*
*Chair, 1996 Hub 4 Evaluation Specification Working Group*

Department of Electrical and Computer Engineering
and School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

## ABSTRACT

This paper reviews and discusses the specification and implementation of the 1996 DARPA Hub 4 Continuous Speech Recognition Evaluation, using speech obtained from a variety of television and radio broadcast news programs. The benchmark test consisted of required partitioned evaluation (PE) and optional unpartitioned evaluation (UE) components. In addition, a variety of acoustical "focus" conditions were identified and separately annotated in the data used for the evaluation. Nine sites participated in the evaluation, with the best system reporting a word error rate of 27.1%. Differences between error rates for the UE and PE components were very small for two out of the three sites reporting results for these two conditions.

## 1. INTRODUCTION

The 1996 DARPA Hub 4 Broadcast News Evaluation was the first integrated effort to transcribe through speech recognition a variety of television and radio broadcast news programs. It represents a continuation and extension of both the common DARPA Continuous Speech Recognition (CSR) evaluations from 1992 through 1995, as well as the pilot evaluation of radio news broadcasts conducted in 1995. This paper reviews the specification and implementation of the 1996 DARPA Hub 4 evaluation, and it briefly discusses some of the issues that the evaluation raises.

## 2. BACKGROUND FOR THE 1996 HUB 4 BROADCAST NEWS EVALUATION

From 1992 through 1995, the common DARPA Continuous Speech Recognition (CSR) evaluation had been based on transcription of read sentences from the *Wall Street Journal*, and more recently on the recognition of sentences read from business articles obtained from a larger set of newspapers [6, 7, 9, 11]. Over the years the format of the CSR evaluation had evolved into a "Hub and Spoke" paradigm which enabled all sites to perform baseline evaluations on a common dataset (the Hub) while at the same time enabling sites to develop systems that addressed specific focal problems in the areas of acoustic or language modeling (the Spokes) [6]. The CSR evaluation task in 1995 was denoted the "Hub 3" task to distinguish its content from earlier core CSR evaluations which used different corpora of speech data.

In 1995 four sites (BBN, CMU, Dragon, and IBM) participated in a new pilot evaluation, called Hub 4, which consisted of transcriptions of the PRI broadcast news program *Marketplace.* These sites generally observed that the task of building a system to transcribe the broadcast news shows was a challenging but very interesting problem. It was also noted that a successful solution to the broadcast-news transcription problem required that sites address most of the facets of speech recognition that were the object of the traditional CSR spoke evaluations. These include spontaneous speech, and adaptation to non-native English speakers, background noise and music, and speech conveyed over the telephone network. In addition, broadcast news sources provided a greater variety of syntax, speaking styles, and acoustical environments than had been a part of the previous CSR evaluations, and they did so in a fashion that was generally perceived to be more "natural". For these reasons, most sites that had participated in either the 1995 CSR evaluation using read speech or the 1995 pilot evaluation of speech from news broadcasts indicated a preference for continuing to use broadcast speech as the domain for DARPA CSR evaluations in the coming years.

## 3. IMPLEMENTATION OF THE EVALUATION SPECIFICATION

For a number of years, the CSR Corpus Coordinating Committee (CCCC) chaired by Francis Kubala in concert with the speech group of the National Institute of Standards and Technology (NIST) had specified and implemented the DARPA CSR evaluation. The CCCC dissolved at the end of 1994, and it was replaced by two *ad hoc* working groups that coordinated the 1995 Hub 3 and Hub 4 evaluations, chaired by Richard Stern and Alex Rudnicky, respectively [10, 11].

It was recognized that it would be difficult for any single individual working on a voluntary and part-time basis to oversee the tasks of specifying the speech and text databases, monitoring their collection and annotation, defining the evaluation specification, and coordinating the efforts of the Linguistic Data Consortium (LDC, which collected and annotated the speech and text data), NIST (which implemented and scored the evaluation), and the participating sites. Because of this difficulty, the various management and coordination tasks were performed in "distributed" fashion by several small working groups with limited responsibility for the following tasks:

- Broadcast news program selection, Long Nguyen, Chair
- Broadcast news recording specification, Matthew Siegler, Chair (later taken over by the LDC)

- Broadcast news transcription specification, Ramesh Gopinath, Chair (later taken over by the LDC)
- Broadcast news language model conditioning, Alex Rudnicky, Chair
- Evaluation specification, Richard Stern, Chair

All of the above groups reported to the Speech Recognition Coordinating Committee (SRCC), which consisted of Charles Wayne (Chair), Richard Schwartz (BBN), Roni Rosenfeld (CMU), Salim Roukos (IBM), and Patti Price (SRI).

| Name | Institution |
|---|---|
| **Core sites:** | |
| Richard Stern (Chair) | Carnegie Mellon |
| Andrej Ljolje,<br>Mike Riley | AT&T Bell Labs |
| Gary Cook,<br>Dan Kershaw | Cambridge University<br>(Connectionist group) |
| Phil Woodland | Cambridge University<br>(HTK group) |
| Lazaros Polymenakos | IBM |
| Jean-Luc Gauvain | LIMSI |
| Chi-Wei Che | Rutgers |
| Ananth Sankar | SRI |
| **Consultants:** | |
| George Doddington | US Government |
| David Graff | LDC |
| David Pallett,<br>Jon Fiscus,<br>John Garofolo | NIST |

**Table 1.** Core members of the Hub 4 Evaluation Specification Working Group.

The core members of the Evaluation Specification Working Group who participated in frequent teleconferences are listed in Table 1. Discussions of the Group and the evolving draft specification were also circulated among a number of additional sites including BBN, BU, Lucent, NYU, OGI, Philips, and SRU. Francis Kubala from BBN, in particular, provided many helpful comments and suggestions. In addition to the obvious task of specifying the evaluation, the Working Group also served as a forum for discussion and resolution of issues that developed during the data collection and annotation process. At several critical junctures George Doddington also provided timely leadership and coordination between LDC and NIST that was needed for the successful completion of the evaluation.

The Hub 4 Evaluation Specification Working Group was charged with the task of developing an evaluation that would improve the basic performance of speaker-independent unlimited-vocabulary recognition systems. It was intended that the combination of recording environments used in broadcast news studios and in the field would "provide impetus to improve core speech recognition capability, to improve the adaptability of recognition systems to new speakers, dialects, and recording environments, and to improve the systems' abilities to cope with the difficult problems of unknown words, spontaneous speech, and unconstrained syntax and semantics."

# 4. SPEECH AND TEXT RESOURCES

## 4.1. Acoustical Databases

The most important new resources developed for the 1996 Hub 4 evaluation were the acoustical and text databases used for training and testing. As described in [3], an acoustical training database for the Hub 4 task was prepared by LDC in coordination with NIST. This database contained approximately fifty hours of speech from eleven news programs produced by ABC, CNN, CSPAN, NPR, and PRI. The data (which were selected from recordings of approximately 130 hours of speech) were recorded, transcribed, and annotated according to acoustical quality and speaker identity (when known) by LDC. Because the recording and annotation of this training corpus and the development of the evaluation specification took place concurrently, the annotations developed by LDC were later passed through a filter developed by NIST to produce a set of labels that corresponded to the focus conditions of the Hub 4 evaluation, as described below.

The speech data used for acoustical training were recorded between May 10 and July 3, 1996. Sites were also permitted to train their systems on any other publicly-available speech data recorded before June 30, 1996, including all previously-released DARPA speech databases.

In addition, LDC also released a developmental test set consisting of approximately three hours of speech from radio and television broadcasts that were recorded between July 10 and July 15, 1996.

The evaluation data consisted of approximately 2.5 hours of broadcast news speech, recorded between September 11 and September 25, 1996. There was a partial overlap of programs in the evaluation test set with the programs that were used for the training set and development test set. The evaluation test set consisted of recordings from a single monophonic channel of audio. Audio segments containing commercials and sports results were excluded from the evaluation.

## 4.2. Text databases

The working group chaired by Alex Rudnicky developed conditioning tools for a large corpus of commercial text transcripts of broadcast new shows similar to those used in the evaluation. These tools and transcripts were prepared and released by the LDC to the participating sites. The text database included about 122 million words in the training set and 19 million words in the developmental test set, drawn from about 110 shows broadcast by ABC, CNN, NPR, and PBS.

Sites were also permitted to make use of any additional publically-available text corpora from shows dating on or before June 30, 1996, including all previously-released DARPA and LDC text databases, but excluding data from shows that were reserved by NIST for the evaluation test set.

There was no standard language model released for the 1996 Hub 4 evaluation, although the language model preparation tools developed by CMU for the 1995 CSR evaluations were widely distributed and used.

# 5. BENCHMARK TEST DESCRIPTION

## 5.1. Acoustical Focus Conditions

The four sites participating in the 1995 pilot evaluation of broadcast news speech all made use of methods that segmented incoming speech into regions exhibiting similar overall acoustical quality [2, 4, 5, 12]. Although the number and nature of the segment classes differed from site to site, sites tended to identify regions of "clean speech", speech in noise, and speech over the telephone, and they would make use of different acoustical models and other recognition parameters for each class. In order to assist sites participating in the 1996 Hub 4 evaluation that did not wish to devote time and effort to the task of segmenting and classifying speech according to acoustical quality, it was decided by the SRCC that the 1996 evaluation would consist of two components: a mandatory *Partitioned Evaluation* (PE) component, in which segment boundaries and labels are provided, and an optional *Unpartitioned Evaluation* (UE) component, in which sites perform the transcription with no side information available that describes acoustical quality.

In order to measure and compare speech recognition accuracy for some of the acoustical conditions of interest to the participating sites, the database was segmented into the following set of *focus conditions* (most of which are reminiscent of objectives of the spokes in earlier CSR evaluations)*:*

- **F0: baseline broadcast speech**, the baseline condition, includes prepared speech recorded in studio conditions
- **F1: spontaneous broadcast speech** includes spontaneous speech recorded in studio conditions
- **F2: speech over telephone channels** includes speech collected under reduced-bandwidth conditions
- **F3: speech in the presence of background music** includes prepared and spontaneous speech at an SNR of 10 to 20 dB, A-weighted
- **F4: speech under degraded acoustical conditions** includes prepared and spontaneous speech degraded by additive noise, environmental noise, or nonlinear distortions, at an SNR of 10 to 20 dB, A-weighted
- **F5: speech from non-native speakers** includes studio-quality intelligible English speech spoken by non-native speakers of American English (including English spoken by natives of the United Kingdom)
- **FX: miscellaneous** includes speech that does not satisfy any of the above conditions, or speech that simultaneously satisfies more than one of the conditions F1 through F5 (such as non-native speech with music in the background)

The development and evaluation test sets were selected to provide "adequate" coverage of the conditions F1 through F5.

It should be noted that the annotation of the speech according to the above criteria proved to be quite difficult, and many of the labelling and segmentation decisions inevitably became matters of personal judgment. As noted above, the LDC actually initiated the process of segmenting and labelling the training data some time before the evaluation specification was completed, using a more verbose set of labels. NIST developed a series of annotation filters that automatically converted the LDC labels into the sets specified for the evaluation.

## 5.2. Other Evaluation Conditions

Sites were permitted to make use of any recognition approach for both the PE and UE evaluations, including unsupervised transcription-mode recognition using multiple decoding passes. Any audio segment in the evaluation test data could be used for adapting any other segment of audio, even from other episodes and shows.

The only side information available for the UE was the locations of endpoints of contiguous blocks of audio, plus the beginnings and endings of commercials and sports results. Sites evaluating on the PE were provided this information plus segment boundaries, story boundaries, and labels according to the acoustical focus conditions conditions F0 through F5. Sites were also provided with information that described which combination of the conditions F1 through F5 were present in each segment classified as FX.

Sites generated decodings that included word time alignments, and word error rate (WER) was calculated by NIST according to the SCLITE scoring package as described in detail in [1].

| Date | Event |
|---|---|
| July 14, 1996 | Evaluation specification approved |
| July 15 | Distribution of 50 hours of acoustic training data |
| July 25 | Distribution of acoustical devtest data |
| July 30 | Distribution of language model text data and tools |
| October 5 | Release of 50 hours of annotated transcripts completed |
| November 11 | Distribution of evaluation test data |
| December 12 | Deadline for submission of core recognition results |
| December 19 | Deadline for submission of contrast results |

**Table 2.** 1996 Hub 4 evaluation schedule.

Table 2 lists some of the milestones in the evaluation schedule. As can be seen in Table 2, there was little time between the release of the complete set of 50 hours of annotated transcripts and the distribution of the evaluation test data, and as a result most sites made less use of the annotated acoustical training data than they had originally intended.

# 6. SYNOPSIS OF EVALUATION RESULTS

Eight sites (BBN, the Cambridge University Connectionist and HTK groups, CMU, IBM, LIMSI, Rutgers, and SRI) submitted speech recognition results for the partitioned evaluation (PE) component of the evaluation. A ninth site, NYU, submitted results obtained by rescoring the speech recognition output of the SRI system. Speech recognition WERs for the PE ranged from 27.1% to 53.8% for the complete test, and from 18.7% to 42.7% on the baseline F0 condition. Detailed results of the evaluation, along with many interesting comments and comparisons, are provided in [8].

Three sites (BBN, CMU, and IBM) submitted results for the optional unpartitioned evaluation (UE) component of the Hub 4 evaluation, with WERs ranging from 31.8% to 38.9% for the complete test. For two of these three sites, the WER for the UE increased by no more than 3% to 5% relative to that site's WER for the PE. (The third site noted that their UE results were adversely affected by a minor programming error.) The ability of sites to develop systems for which the UE WER is comparable to the PE WER suggests that it is not extremely difficult for recognition systems to provide appropriate segmentation and blind labeling of segments of incoming utterances.

One of the four shows used for the 1996 Hub 4 evaluation was the PRI *Marketplace* program that had also been used in 1995. BBN and CMU were the only sites that participated in both the 1995 pilot Hub 4 evaluation and the 1996 Hub 4 UE evaluation. The relative WER for the *Marketplace* portion of the 1996 UE component for these sites decreased by about 25% to 30% relative to the WER obtained for that show in the 1995 evaluation.

## 7. DISCUSSION AND COMMENTS

The enthusiastic participation by nine sites in the 1996 Hub 4 evaluation and the decrease in error rates obtained by the participants are both encouraging. The experience in 1996 confirmed the expectations that the broadcast news task is both more challenging and more relevant than the read-speech tasks used in previous CSR evaluations.

As noted in Sec. 5, the time available for detailed study and analysis of the training database collected for the evaluation was limited. Since the Hub 4 evaluation protocol in 1997 will be essentially the same as in 1996, and since an additional 50 hours of annotated broadcast news training data was distributed by LDC in early 1997, the 1997 evaluation will provide a good opportunity to assess the extent to which 1996 performance was limited by the lack of opportunity to explore and exploit the nature of the task as revealed by the training data.

Finally, we note that while the distribution of management tasks among a number of individuals worked reasonably well for some aspects of the evaluation, centralized responsibility and oversight can be extremely helpful in resolving unforseen problems in a timely fashion.

## ACKNOWLEDGEMENTS

Preparation of this report was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The author thanks all the members of the Hub 4 Evaluation Spcficiation Working Group, and especially Charles Wayne for his support and advice, and George Doddington, Dave Graff, and Dave Pallett for their contributions to the success of this evaluation.

## REFERENCES

1. Garofolo, J. S., Fiscus, J. G. , and FIsher, W. M., "Design and Preparation of the 1996 Hub-4 Broadcast News Benchmark Test Corpora", these Proceedings.

2. Gopalakrishnan, P. S., Gopinath, R., Maes, S., Padmanabhan, M., Polymenakos, L. C., Printz, H., and Franz, M., "Transcription of Radio Broadcast News with the IBM Large Vocabulary Speech Recognition System", *Proceedings of the ARPA Speech Recognition Workshop,* Morgan Kaufmann Publishers, pp. 72-76, February, 1996.

3. Graff, D., Wu, Z., MacIntyre, R., and Liberman, M., "The 1996 Broadcast News Speech and Language-Model Corpus", these Proceedings.

4. Jain, U., Siegler, M. A., Doh, S.-J., Gouvêa, E., Huerta, J., Moreno, P. J., Raj, B., and Stern, R. M., "Recognition of Continuous Broadcast News with Multiple Unknown Speakers and Environments", *Proceedings of the ARPA Speech Recognition Workshop,* Morgan Kaufmann Publishers, pp. 61-66, February, 1996.

5. Kubala, F., Anastasakos, T., Jin, H., Makhoul, J., Nguyen, L., Schwartz, R., and Yuan, N., "Toward Automatic Recognition of Broadcast News", *Proceedings of the ARPA Speech Recognition Workshop,* Morgan Kaufmann Publishers, pp. 55-60, February, 1996.

6. Kubala, F., Bellegarda, J., Cohen, J., Pallett, D., Paul, D., Phillips, M., Rajasekaran, R., Richardson, F., Riley, M., Rosenfeld, R., Roth, B., and Weintraub, M., "The Hub and Spoke Paradigm for CSR Evaluation", *Proceedings of the Spoken Language Technology Workshop*, Morgan Kaufmann Publishers, pp. 9-14, March, 1994.

7. Kubala, F., "Design of the 1994 CSR Benchmark Tests", *Proceedings of the ARPA Spoken Language Systems Technology Workshop*, Morgan Kaufmann Publishers, pp. 41-46, January, 1995.

8. Pallett, D. S., and Fiscus, J. G., "1996 Preliminary Broadcast News Benchmark Tests", these Proceedings.

9. Paul, D., Baker, J., "The Design for the Wall Street Journal-Based CSR Corpus", *Proceedings of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Publishers, pp. 357-362, February, 1992.

10. Rudnicky, A. I., "Hub 4: Business Broadcast News", *Proceedings of the ARPA Speech Recognition Workshop,* Morgan Kaufmann Publishers, pp. 8-11, February, 1996.

11. Stern, R. M., "Specification of the 1995 ARPA Hub 3 Evaluation: Unlimited Vocabulary NAB News Baseline", *Proceedings of the ARPA Speech Recognition Workshop,* Morgan Kaufmann Publishers, pp. 5-7, February, 1996.

12. Wegmann, S., Gillick, L., Orloff, J., Peskin, B., Roth, R., van Mulbregt, P., and Wald, D., "Marketplace Recognition using Dragon's Continuous Speech Recognition System", *Proceedings of the ARPA Speech Recognition Workshop,* Morgan Kaufmann Publishers, pp. 67-71, February, 1996.